**The Thesis Committee for Lisi Wang**
**Certifies that this is the approved version of the following thesis:**

# Investigating Transfer-Appropriate Processing as a Theoretical Account for the Testing Effect

**APPROVED BY**
**SUPERVISING COMMITTEE:**

**Supervisor:**

Andrew Butler

Diane Schallert

# Investigating Transfer-Appropriate Processing as Theoretical Account for the Testing Effect

**by**

**Lisi Wang, B.S.**

**Thesis**

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

**Master of Arts**

**The University of Texas at Austin**

**August 2016**

## Abstract

### Investigating Transfer-Appropriate Processing as a Theoretical Account for the Testing Effect

Lisi Wang, M.A.

The University of Texas at Austin, 2016

Supervisor: Andrew Butler

Numerous theories have been put forth to explain the mnemonic benefits of retrieval practice relative to restudying (the testing effect). Among these accounts is the concept of transfer-appropriate processing, which is commonly invoked but rarely directly tested. Following up on research by Peterson and Mulligan (2013), the type of intervening task (restudy vs. test), the type of processing (item-specific vs. relational) during the intervening task, and the type of processing during the final test were manupulated in a between-subject design. Participants studied rhyming cue-target word pairs, and then either restudied the pairs or took a test on the target words. In these learning activities, cues were either randomly presented (item-specific processing) or grouped by semantic categories (relational processing). In the delayed final test, participants were assessed by cued recall (item-specific processing) or free recall (relational processing). The pattern of results supports the transfer-appropriate processing

account when the final test was free recall, but the opposite pattern was observed in cued recall.

## Table of Contents

# List of Tables

# List of Figures

# Chapter 1: Introduction

Between initial learning and taking a final test, taking a practice test as opposed to restudying the same materials can greatly improve performance on the final test. This is the testing effect. The action of taking the practice test is termed "retrieval practice" (Roediger & Karpicke, 2006b; Roediger & Butler, 2011). This finding has been replicated exhaustively with materials ranging in complexity from word lists in laboratory experiments to complex learning materials in educational settings (Roediger & Karpicke, 2006a). Given these benefits, retrieval practice has been suggested as an effective learning strategy (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). However, theoretical accounts of the testing effect still remain equivocal (Karpicke, Lehman, & Aue, 2014). This introduction aims to review some major theories that explain why retrieval practice often produces superior retention relative to restudy and how different types of retrieval practice differ from each other. Some theories are more descriptive, while others focus on identifying a potential mechanism. One important thing to keep in mind is that although divergent, these theories are not mutually exclusive. An experiment predominantly testing the theory "transfer-appropriate processing" will be described and predicted results based on other theories will be made.

## THEORETICAL ACCOUNTS

### Retrieval Difficulty

Bjork (1975) proposed that it is the difficulty of retrieval practice that alters memory and leads to higher performance on the final test. Here, retrieval difficulty is roughly equated to depth of encoding (Craik & Lockhart, 1972; Craik & Tulving, 1975). Bjork argued that in terms of retention, processes during retrieval practice are at least as crucial as processes during initial encoding. The wide-held belief of level of processing during encoding is therefore made applicable to retrieval practice. For example, when trying to remember a list of words, the first few words can be processed in a more meaningful and elaborative fashion, whereas the last few words are usually subject to rote memorization due to interference. In an immediate test, the first few words and the last few words are better remembered than words in the middle of the list (the positive recency effect). After a long delay, memory of the first few words persists, while the last few words are remembered poorly (the negative recency effect). Elaborative rehearsal and retrieval practice from long-term memory represent a deeper level of processing than rote rehearsal and retrieval practice from short-term memory, which leads to the first few words being retained better than the last few words in the long run. Similar reasoning applies to semantic vs. phonological processing (Whitten, 1978) and longer vs. shorter processing time (Gardiner, Craik, & Bleasdale, 1973).

In this fundamental paper (Bjork, 1975), cognitive mechanisms underlying different levels of processing are not explained in details. It is implied that retrieval

practice acts as a memory modifier that reconstructs memory representations of learned information. Deeper levels of processing triggered by difficult retrieval tasks can strengthen representations and multiply retrieval routes to previously learned items.

**Retrieval Strength**

The strengthen representations part mentioned above is later elaborated by Bjork and Bjork (1992) in the new theory of disuse. This is an overarching theory aiming at describing "the architecture of human memory" including retrieval and forgetting, hence "disuse". The new theory of disuse proposes that strength of memory representations can be seen as storage strength (quality of representations) and retrieval strength (accessibility of representations in the presence of retrieval cues). Storage strength and retrieval strength are functionally independent and successful retrieval is predominantly dependent on retrieval strength. Upon successful retrieval of previously learned items, storage strength and retrieval strength of the items are increased. Therefore, being a retrieval activity itself makes taking a test superior to restudying. Another claim relevant to the testing effect is that difficult tests can facilitate more enduring learning than easy tests. Although successful retrieval builds up retrieval strength, its growth rate decreases as current retrieval strength increases. Consequently, successfully retrieving items with lower retrieval strength to begin with can lead to larger increment in retrieval strength. Bjork (1994) recommended that this "desirable difficulty" be implemented in real-life learning.

The new theory of disuse expands retrieval difficulty into a more concrete description. However, this development does not change the qualitative nature of both accounts. Verbal descriptions without quantitative indicators do make both accounts speculative. For example, difficulty of retrieval does not consist a scale per se, but is defined in relation to easy learning activities. In fact, interpreting retrieval difficulty in the same context can be ambivalent. In several studies, retrieval difficulty was artificially increased by extending processing time and it was found that longer processing time yielded higher performance on the final test (e.g., Pyc & Rawson, 2009). However, it is equally possible that reduced processing time can make learning tasks more difficult (Karpicke et al., 2014), thus changing the direction of the conclusion.

**Bifurcation**

Put forth by Kornell, Bjork and Garcia (2011), this model is in line with the strength theories, but provides a visual representation of how memory strength changes through learning activities and time. One basic assumption is that strength of representations of all studied items can be quantified on a continuum that is approximately normally distributed. Upon intervening test, items high on the continuum are successfully retrieved, while the others are not. This retrieval attempt then bifurcates strength of all items: only those items retrieved will increase in strength. In comparison, all items studied for a second time will increase in strength, but to a lesser extent. Another basic assumption is that through time, strength of all items "decay" at a roughly equal rate. Given that final test performance depends on its inherent retrieval threshold,

only items with strength above a certain value can be successfully retrieved. Due to bifurcation, more tested items would remain above this threshold than restudied items. Providing feedback after a test can minimize bifurcation and lead to an even larger advantage of test over restudy. This is because in addition to being re-exposed to the items, corrective feedback can function as a metacognitive monitor of progress and motivate further learning (Roediger & Karpicke, 2006b).

The assumptions mentioned above are largely unknown and untestable. It is unlikely that strength of items can be plotted as a normal distribution or strength of items decline indiscriminately after different intervening tasks. However, this model is made unique by emphasizing final test requirements. Retrieval threshold of the final test determines the amount of items that can be retrieved, as well as the extent to which test performance surpasses restudy performance. For instance, free recall tests are more demanding than cued recall tests. If the bifurcation model is correct and test does strengthen retrieved items to a greater extent than restudy, the advantage of test over restudy would appear more pronounced in a free recall than a cued recall final test. This is exactly what Halamish and Bjork (2011) found when comparing cued recall and study as intervening tasks.

**Encoding Variability**

The following three theories all deal with cognitive mechanisms that multiply retrieval routes from cues to targets to improve memory and learning performance. Encoding variability is the earliest and most fundamental of them. It is mostly relevant to

the spacing effect in retrieval practice (distributed practice produces better long-term retention than massed practice) (Melton, 1970). In paired-associate verbal learning, Martin (1968) proposed that stimulus meaningfulness is inversely related to variability of encoding events and perceived stimulus-response associations. Therefore, multiple encounters with a less meaningful stimulus can form various stimulus-response associations and exhibit more positive transfer to different testing environments, whereas responses to more meaningful stimuli are stable and exhibit more negative transfer. Melton (1970) situated this model in a more concrete setting to explain the spacing effect in free recall. Variability in encoding informs us that the reason why test is better than restudy is that test introduces more variability than restudy. The encoding variability account also explains why some types of tests are better than others. In the context of the testing effect or retention in general, the more retrieval routes from cues to targets exist, the more likely targets can be accessed. Such cues can be very general in scope and include internal and external context, semantic information and structural organization of cue-target associations (Glenberg, 1979).

**Elaborative Retrieval**

Elaborative retrieval and encoding variability differ in their focuses when it comes to describing how retrieval routes are formed. In encoding variability, perceptions of retrieval cues vary from encounters. In elaborative retrieval, what makes learning experiences different is whether retrieval routes between cues and targets are elaborative or not. Inspired by spreading activation of semantic networks (Collins & Quillian, 1972),

Carpenter (2009) hypothesized that when presented with retrieval cues and asked to recall target items, people engage in an active searching process. This searching process allows mediators semantically related to the cues to be activated thus elaborate retrieval routes to the targets. Associations between cues and targets can be enhanced by test, because targets are available in restudy and there is no need to search for them. Similarly, difficult intervening tests can lead to better learning than easy tests. This is because there is more elaborative searching involved and more retrieval routes formed in difficult tests.

A major limitation is that the body of research behind this hypothesis is largely correlational. When this hypothesis was tested by Carpenter (2011), participants were asked to learn a number of semantically related cue-target words (e.g., *Mother-Child*) in the initial encoding phase. Those in the test condition were prompted by a cued recall of the target words (e.g., *Mother-*_____), whereas those in the restudy condition simply read the associations again. In the final phase, both groups of participants were tested by recognition and cued recall. It was found that the test group falsely recognized more mediators semantically related to the cue words (e.g., *Father*) than the restudy group. However, this pattern was not seen on words unrelated to the cue words (e.g., *Rabbit*). Not surprisingly, participants in the test condition were more likely to recall the target words when cued by the mediators (e.g., *Father-*_____) than by words related to the targets, but not the cues (e.g., *Birth-*_____). Carpenter (2011) stated that it is the semantic mediators that are activated to enable elaborative retrieval paths to form cue-target associations. Since responses to the semantically related words were measured post hoc, it is difficult to interpret what actually happened in real time. It is equally convincing that

semantically related words could be activated after the cue-target associations are made. Associating targets to cues may not depend on semantic mediators per se, instead, direct retrieval routes could be formed from cues to targets and other semantically related words would become activated as a byproduct.

**Episodic Context**

In a similar vein as enabling encoding variability and multiplying retrieval routes from cues to targets, the episodic context account specifies that contextual cues are updated to include the past and present contexts, thus the aggregate of contextual cues uniquely manifest targets. This recent theory advanced by Karpicke and colleagues (2014) precisely outlines the underlying cognitive processes that may eventually transform episodic learning events to semantic information devoid of contexts. After a learning event, information about the event is incompletely stored as item features and context features. When retrieving previously learned information, people use available contextual cues to search for item features and reconstruct item features. Due to constantly shifting contexts from past to present, people need to reinstate previous contextual cues (context reinstatement). Context representation of this event therefore includes past and present contexts (context updating). Because there are fewer items associated with the updated context representation, this allows people to restrict their search set when asked to retrieve the event again (restriction of search set). After multiple encounters with the to-be learned item, items unique to the context representation can even become decontextualized.

The episodic context account is able to identify cognitive processes involved in the testing effect and related phenomena. Because fewer contextual cues are provided in test than in restudy, reinstatement of past context is mandatory. Context reinstatement is one of the key processes that drive learning. The way difficult tests are compared with easy tests can be made similarly. For example, it was found that an intervening test consisting of weak semantic cues (e.g., *Basket-*_____) produced higher free recall performance than strong semantic cues (e.g., *Toast-*_____) (Carpenter, 2009). Carpenter originally proposed that weak semantic cues can activate more semantic elaboration and generate more retrieval routes to target words (e.g., *Bread*). The episodic context account offers an alternative explanation: more context reinstatement and update are performed in the presence of weak cues, which leads to more directed search of the target word.

**Transfer-Appropriate Processing**

Other than bifurcation, transfer-appropriate processing (TAP) is another theory that emphasizes final test requirements. TAP is initially proposed to expand the level of processing account of encoding to processing during encoding and retrieval (Craik & Lockhart, 1972; Craik & Tulving, 1975). The central argument is that all learning processes are inherently equal. Judgments of values associated with these learning processes can only be made by comparing them with final tests processes (Morris, Bransford, & Franks, 1977). Final test performance is high when cognitive processes in initial learning and final test overlap and optimized when they are identical. That is, the processes "transfer" from encoding to retrieval. This theory bears similarities to the

encoding specificity principle (Tulving & Thomson, 1973), which states that it is the contextual consistency between encoding and retrieval that are predicting performance on the final test. In comparison, TAP focuses more on the cognitive processes activated during both initial learning and final test.

When TAP is applied to explaining the testing effect, test is superior to restudy simply because the nature of a final test is always some form of testing. An intervening test would elicit cognitive processes potentially more useful to the final test than what a restudy session would do. No specific prediction of some types of tests being superior is mentioned in the study by Morris and colleagues (1977), but Experiment 1 procedures and results can be informative. Participants were involved in an incidental learning task consisting semantic encoding (e.g., *The TRAIN has a silver engine*.) and rhyme encoding (e.g., *EAGLE rhymes with legal*.). At the end, they were asked to perform semantic recognition of original words or rhyme recognition of rhyming words. It was found that items semantically encoded were better recognized than those phonologically encoded in the semantic retrieval task, whereas the opposite was observed in the rhyme retrieval task. This pattern no doubt indicates performance is higher when encoding and retrieval processing are matched than mismatched, exactly what TAP would predict. In addition to this interaction, main effects of semantic encoding and semantic retrieval were statistically significant. As mentioned earlier, theories accounting for the testing effect are not mutually exclusive. It is plausible that TAP and level of processing/retrieval difficulty accounts coexist. In the context of retrieval practice, difficult intervening tasks can still promote overall higher final performance than easy tasks, but their relative

contributions need to be evaluated with reference to requirements or difficulties of the final tests.

TAP qualifies cognitive processes involved in encoding by emphasizing that processing consistency between encoding and retrieval is the key to optimizing performance. TAP sets limitations to a number of other theories targeting encoding processing only and provides a processing framework for studying encoding and retrieval. However, the TAP account has some inherent limitations. It has been criticized that explaining studies with TAP would essentially lead to a circular argument. On the one hand, TAP offers a general description of processing without specifying actual cognitive processes. On the other hand, it only becomes accountable after the final test is complete. Consequently, discovering precise mechanisms involved in encoding and retrieval is not encouraged, simply because precise mechanisms do not matter as long as they are consistent throughout the experimental design.

These seven theories, together with their proponents, focus on description or mechanism, key information, and related retrieval practice phenomena they best explain are summarized in Table 1.

## STUDIES INVESTIGATING THE TAP ACCOUNT FOR THE TESTING EFFECT

After reviewing the major theories of the testing effect, a general trend can be observed from explaining changes in cue-target association strength to illustrating complex cognitive processes in retrieval practice. It is also shown that some theories are more suited to accounting for difficulty of tests (e.g., retrieval difficulty, retrieval

strength, elaborative retrieval, episodic context), while others make prediction based on completeness of tests: new elements integrated into representation from testing (e.g., encoding variability, episodic context). TAP is especially interesting, because it provides a framework that includes the nature of the final test. It states that cognitive processes during initial learning or the intervening test are not the sole determinant of performance, but their effects are conditioned on cognitive processes the final test calls for. Although TAP may not support our investigation into exact mechanisms, studying different types of cognitive processes interacting between encoding and retrieval is still permitted. TAP also enables us to compare and contrast the nature of cognitive processes involved in encoding and retrieval. Given the importance, this section seeks to review some studies employing the encoding-retrieval paradigm (Tulving, 1983, p. 219), that is, studies that factorially crossed processing conditions between initial learning or the intervening task and the final test. TAP and the encoding specificity principle were advocated almost 40 years ago, but only a limited amount of research studied the testing effect in the using the recommended design. As expected, not all the studies are suggestive of TAP. Advantages and limitations of these studies will be discussed, which will lead to the current experimental design.

**Semantic and Orthographic Processing**

A study by Veltre, Cho & Neely (2014) nicely demonstrated the TAP account in retreival practice. The initial learning phase consisted of learning lists of words (e.g., *ABOVE*). In the next phase, participants encountered a recall test with target words cued

by semantic cues (e.g., *BELOW*) and orthographic cues (e.g., *A_O_E*). The final recall was cued by cue words that would elicit the same types of processing as the test and those that would elicit different types of processing. Assuming the word "ABOVE" was recalled under semantic processing in the intervening test, examples of the same type of final cues would be "BELOW" (identical) and "BEYOND" (same type). An example of a different type would be "A_O_E" (different type). Together, there were six conditions.

An ANOVA with types of final cues and whether intervening task cues and final cues were the same or different showed that the match between cues or processing was statistically significant, supporting the TAP account. Looking at performance with semantic final cues and orthographic final cues separately, the same types of processing during the intervening test on average led to higher performance on the final test than different types of processing, but this was only statistically significant when the final cues were semantic. As a side note, this study also compared test with study as intervening tasks. Reviewing target words in a random order (the intervening study) functioned as a generic control condition. Because the control condition did not prompt semantic or orthographic processing same as the test conditions, differences between test and study undergoing the same processing could not be made.

**Semantic and Phonological Processing**

McDaniel and Masson (1985) conducted a study that showed partial support for TAP, between initial learning and final test. The main purpose of their study was to investigate how variations in learning events affect performance. However, there is

information relevant to TAP. Participants initially learned a list of words (e.g., *HAWK*) by making semantic (e.g., *How well does the word categorize with BLACKBIRD?*) or phonological (e.g., *How well does the word rhyme with TALK?*) judgments. The final test asked participants to recall the words from similar but different semantic and phonological cues. Details of the intervening tasks are not important for the present purpose, because data obtained under different processing were combined. Intervening test (Experiment 1) was compared with intervening study (Experiment 3).

Main effects of initial semantic encoding and final phonological recall were found. Also evident was an interaction between initial and final processing. However, superiority of consistent processing between initial encoding and final recall was only confirmed when under phonological processing. McDaniel and Masson (1985) explained that this was potentially due to rhyming cues having less potential associations than semantic cues. That is, the rhyming final test was easier than the semantic final test, which led participants to perform better in response to rhyming cues, even though the target words were initially encoded semantically. The limitation in the study by Veltre and colleagues (2014) was avoided by McDaniel and Masson (1985), because intervening test and study underwent the same types of processing. Nevertheless, these potentially fruitful analyses were merged into test versus study. This made the study less informative to the current experiment, because the ultimate goal is to examine cognitive processes during the intervening task and the final test. Another change the current experiment needs to make is to employ intentional learning as opposed to incidental learning, which can make it more comparable to goal-directed learning in educational settings.

**Relational and Item-Specific Processing**

Different mechanisms underlie cognitive processes in semantic, phonological, and orthographic modalities. Likewise, relational processing and item-specific processing differ in terms of mechanisms. Relational processing helps abstract similarities among items, whereas item-specific processing highlights distinctive features of each item (Einstein & Hunt, 1980; Hunt & Einstein, 1981). This statement is based on a functional dissociation where retention was affected by relational and item-specific processing jointly but differently (Einstein & Hunt, 1980). A relational task (taxonomic organization) improved category clustering in free recall to a large extent, but had little to do with increasing recognition scores or decreasing false alarm rates. In contrast, the opposite pattern was obtained with a task item-specific in nature (pleasantness rating). If this distinction between relational and item-specific processing holds true, TAP should be evident in studies utilizing them during the intervening task and final test, which is what Peterson and Mulligan (2013) found.

Peterson and Mulligan (2013) conducted a series of experiments where they made subtle adjustments to the designs and changed the direction of the testing effect. Experiment 1 demonstrates a surprising finding of a negative testing effect. Participants learned a list of rhyming cue-target associations (e.g., *Moon-Spoon*) in the initial encoding. The target words were from six distinct semantic categories (e.g., *Spoon* is an exemplar of the kitchen utensil category). Then, participants either generated the cue words from the given target words (test), or read the cue-target associations again (restudy). Note that partial and full pairs were presented in the same semantic category

blocks. Final performance was measured by free recall. As indicated by percentage recalled and category clustering, test instead of restudy actually disabled performance on free recall. Peterson and Mulligan (2013) further delineated relational and item-specific processing into inter-item relational processing of categories, intra-item relational processing of cue-target pairs, and item-specific processing of targets. They explained that generating targets from cues enhanced item-specific and intra-item relational processing. However, since test was too cognitively demanding than restudy, those in the test condition were less attentive to inter-item relational information than those in the restudy condition, which weakened their performance on the final free recall test.

When processes involved in the intervening test and processes required by the final test were matched, the negative testing effect reverted to positive testing effects. This was achieved by changing the final test from free recall to cued recall (Experiment 2) and presenting the intervening task in a random order, instead of in category blocks (Experiment 3). The former case reduced final test requirement on inter-item relational processing, while the latter case limited inter-item relational processing during the intervening task. The combination of these three studies is congruous with the TAP account. Whether test appears more advantageous than restudy to raise final performance depends on the extent to which intervening test processing and final test processing are compatible with each other. Interestingly, although concentrating on consistency in processing, this study was not posited in a design factorially manipulating processes in the intervening task and the final test. A factorial design could reduce type I error and make results supporting TAP easy to identify, because TAP essentially implies an

interaction between the intervening task and the final task. Another reason why this study should to be replicated is that conflicting results regarding the negative testing effect have been found (Rawson, Wissman, & Vaughn, 2015), which will be covered in details later.

**Format of Test**

Free recall and recognition may entail cognitive processes that can be dissociated. According to Jacoby (1991), recollection and familiarity judgment are distinct processes. This is because the former is an intentional process, while the latter is an automatic process. In memory tests, the two types can be measured by free recall and recognition and respectively. If these two types of test are truly dissociable, they are expected to show the TAP in an encoding-retrieval paradigm. For instance, a recognition intervening test can improve a recognition final test to a greater extent than a recall intervening test, but when the final test is recall, a recognition intervening test is no longer at advantage. However, this expectation was not met according to the studies to be analyzed.

Glover (1989) crossed free recall, cued recall, and recognition between intervening test and final test and found an overall effect of free recall as an intervening test on all forms of final tests (Experiment 4). Specifically, participants were asked to read a prose passage in the initial learning phase. In the intervening task phase, participants were either tested by free recall, cued recall, and recognition, or dismissed as baseline control. All participants conducted free recall, cued recall, or recognition in the final test. When each final test format was analyzed separately, intervening free recall appeared to be universally more beneficial than cued recall and recognition to the final

test. This led Glover (1989) to conclude that it is the completeness of retrieval practice event that is driving the final performance. Retrieval practice experiences elicited by free recall, cued recall, and recognition decrease in completeness, because cued recall and recognition increase in the amount of retrieval "entry points" already provided.

There are certain limitations associated with this study by Glover (1989). Processing time was a potential confounding variable with completeness of the intervening test. There was no time constraint on the intervening test, thus it could be possible that participants who took the free recall intervening test spent a longer time than those in the other conditions. This may rule out the completeness of retrieval event explanation. Another problem was that counterintuitively, in the final test phase, free recall performance was higher than cued recall performance and both of them were superior to recognition performance. Because the final test was in the same format as the intervening test, this may mean that free recall, cued recall, and recognition tests employed in the intervening test increased in difficulty. In fact, the free recall test had the least number of idea units assessed. Performance on the intervening test would provide insight into this postulation, but no data were available from this study.

The general effect of free recall on all final tests continued to exist when Glover's (1989) design was slightly modified by Carpenter and DeLosh (2006, Experiment 1). Carpenter and DeLosh (2006) situated their study in a more controlled setting by adding a restudy condition as control and equating participants in all conditions in their exposure time to the learning material. Another modification was simplifying the learning material using word lists, instead of prose passages. For each type of final test, free recall practice

produced more improvement than cued recall and cued recall yielded larger improvement than recognition. Some parts of this pattern were not statistically significant, but were different in values for the most part. Carpenter and DeLosh (2006) advocated for an elaborative retrieval explanation, due to recognition, cued recall, and free recall being facilitated by less and less available cues during the intervening task, which demanded more and more elaborative processing. Nevertheless, the tests having unusual difficulty levels problem faced by Glover (1989) was not completely avoided. This time in the final test, free recall was more difficult than recognition, but still seemed easier than cued recall. Intervening test performance revealed that participants had similar performance in free recall and cued recall.

Kang, McDaniel, and Roediger (2007) observed that advantages of free recall over cued recall and recognition remained using a similar paradigm, disproving the TAP account. This study was designed to mimic retrieval practice in real-life learning. Prose passages were provided in the first phase, followed by short answer questions, multiple-choice questions, lists of statements to be reread, and irrelevant filler tasks. Only short answer and multiple-choice were repeated in the third phase which made up the final test. The researchers precisely controlled for difficulty by testing the same idea units in different formats. Exposure to materials was not manipulated by equating processing time, but by providing correct answer feedback after the intervening test. Results were not surprising given what we discussed earlier. For both short answer and multiple-choice final tests, rereading statements, taking a multiple-choice test, and taking a short answer test were increasingly beneficial. Again, some comparisons were not significant

statistically, but were different numerically. Kang and colleagues (2007) concluded the article supporting a testing effect account featuring retrieval effort and difficulty. Short answer is more difficult and naturally requires more effort, because no hint is provided for reconstruction. As a side note, their manipulation of difficulty was successful as indicated by short answer performance being lower than multiple-choice performance for both intervening and final tests.

One thing these studies that do not support the TAP account have in common is that they made use of different formats of testing: free recall, cued recall, and recognition. As pointed out by Jacoby (1991), free recall and recognition are usually not process pure in reality. In addition, in terms of relational and item-specific processing, free recall is substantially reliant on relational information, but still needs a fair share of item-specific information. In comparison, recognition is more dependent on item-specific processing (Einstein & Hunt, 1980; Hunt & Einstein, 1981). The reason why free recall contributed more than cued recall and recognition to final tests, including those assessed by cued recall and recognition may be attributable to the part of item-specific processing free recall facilitated in a unique way. This study, therefore, aims to tackle this problem by learning activities that are more process-pure.

## THE PRESENT STUDY

The present study has two goals. The first goal is to investigate whether TAP is a valid theoretical account for the testing effect. The design by Peterson and Mulligan (2013) is expanded according to the encoding-retrieval paradigm. Their learning

materials continue to be used and are presented in ways to trigger relational (cue-target associations presented in semantic category blocks) or item-specific processing (cue-target associations presented randomly) in the intervening task. Relational processing and item-specific processing during the final test are tapped by free recall and cued recall respectively. Note that item-specific processing in this study actually refers to a combination of item-specific and intra-item relational processing as appeared in Peterson and Mulligan's (2013) study. This conversion is consistently used in this study for convenience. More importantly, although Peterson and Mulligan (2013) differentiated item-specific and intra-item relational processing, there measures were not sensitive enough to distinguish between these two types of processing

A byproduct of this study is to compare test and restudy under the same processing conditions. In total, there are three variables: intervening task, intervening task processing, and final test processing. Some other design features include controlling for test difficulty and exposure to material. Relational and item-specific processing in the intervening task and the final test essentially ask participants to attend to the same items. In the intervening task, the same amount of processing time is allocated to all conditions. Furthermore, correct answer feedback is provided after participants respond to the intervening test, thus further assure exposure to items being equivalent across all conditions.

The TAP account predicts an interaction between intervening task processing and final test processing. That is, final test performance is expected to be higher when cognitive processes involved in the intervening task and the final test are matched than

mismatched. Results supporting other theories may also be present, as testing effect theories are not mutually exclusive. For example, in the studies by Morris and colleagues (1977) and McDaniel and Masson (1985), while a deeper level of processing during encoding led to higher performance overall, this level of processing effect coexisted with the consistency effect. In fact, some theories that are different on the surface are complementary in the core concept they are trying to explain, thus make similar predictions. For instance, the three studies discussed in the "contrasting evidence" section had almost identical results, but justified three different theories including completeness of retrieval, elaborative retrieval, and retrieval difficulty and effort, indicating they are only different in perspectives.

In general, test is predicted to yield higher final test performance. This is because test with less information provided is more difficult than restudy (the retrieval difficulty, retrieval strength, and elaborative retrieval accounts). This can also be due to more incidents of context reinstatement in the test condition where retrieval routes from cues to targets are only partially presented (the episodic context account). Retrieval difficulty and retrieval strength theories expect the item-specific intervening task to be more difficult than the relational intervening task. To recap, the item-specific task presents cue-target pairs in random order, whereas the relational task blocks the pairs in categories. Elaborative retrieval, encoding variability, and episodic context theories predict that the relational task is better than the item-specific task for learning as measured by performance on the final test. Elaborative retrieval requires information to be elaborated to be semantic in nature. Blocking target items according to their taxonomic categories is

more semantic than random presentation. Structural information is a type of cue critical to increasing cue-target retrieval routes (Glenberg, 1979), thus encoding variability supports an overall higher performance in the final test as a result of relational processing in the intervening task. The only theory other than TAP that informs us about the final test is bifurcation. Free recall has a higher retrieval threshold than cued recall. Consequently, performance on the final test is predicted to be higher for cued recall than free recall.

The second goal is to replicate the findings by Peterson and Mulligan (2013). When Rawson and colleagues (2015, Experiment 4b) attempted to replicate the design by Peterson and Mulligan (2013, Experiment 1) by comparing test via typing and restudy via reading, a positive testing effect appeared. Peterson and Mulligan (2013) argued that the negative testing effect they found was due to participants in the test condition directing too much attention to processing item-specific and intra-item relational information, which compromised their available cognitive resources for inter-item relational processing. However, as recorded by Rawson and colleagues (2015), more inter-item relational processing was facilitated in the test condition, as demonstrated by higher values in the relational processing measures category clustering. When Rawson and colleagues (2015) later realized that Peterson and Mulligan (2013) asked their participants in the restudy condition to actively produce the cue-target pairs, they changed their design to comparing test via typing and restudy via typing (Experiment 4a). To make the situation even more puzzling, overt study via typing led to worse inter-item relational processing and overall performance than covert study via reading.

It is important to point out that both sets of results are not against TAP or the idea that free recall requires a fair share of both relational and item-specific processing. The key difference is whether inter-item relational processing is suppressed or encouraged during the intervening test. The researchers themselves were uncertain how the same design led to disagreeing results and called for more follow-up studies to resolve these contradictory findings. The part of the factorial design in the current study replicating Experiment 1 of Peterson and Mulligan (2013) and Experiment 4 of Rawson and colleagues (2015), that is, comparing test and study when the intervening task organizes cue-target pairs in category blocks and the final test is free recall is of specific importance.

# Chapter 2: Methods

## PARTICIPANTS

Participants were 156 students (48 men and 108 women). They were recruited from a Department of Educational Psychology Subject Pool. Participants took part on a voluntary basis in exchange for course credits. Five participants were excluded from the dataset at the end of the data acquisition period, because three of them did not follow instructions properly, one participant had very limited English proficiency, and one participant was severely visually impaired.

## MATERIALS

Given that one goal of this study was to replicate the findings by Peterson and Mulligan (2013, Experiment 1), the same materials (see Appendix A) were used. In total, there were 36 cue-target word pairs, with six target words in each of the six semantic categories. The target words were taken from the semantic categories of Van Overschelde, Rawson, and Dunlosky (2004). Cue words rhymed with corresponding target words, but were not in any of the six semantic categories. The cue word "*Wear*" was changed to "*Fair*", because there were two cue words that rhymed with each other and had the same spelling,

## DESIGN

This study employed a 2 (intervening task: test, restudy) × 2 (intervening task processing: relational, item-specific) × 2 (final test processing: relational, item-specific)

between-subject design. Randomly presented cue-target word pairs assessed by cued recall, combined with word pairs grouped by semantic categories assessed by free recall denoted matched processing between intervening task and final test. Randomly presented word pairs measured by free recall, together with word pairs grouped by semantic categories measured by cued recall represented mismatched processing between intervening task and final test.

**PROCEDURE**

This experiment was computer based. Participants were randomly assigned to one of the eight conditions. Each condition consisted of three phases (see *Figure 1*). In Phase I, the 36 target words appeared on the computer screen one at a time in random order. Each word was displayed for 4 seconds. Participants were instructed to remember all the target words for future testing.

Phase I and Phase II were separated by a "spot the difference" game as a distractor task, to eliminate the effects of short-term memory on the following recall tests. The distractor task contained three pairs of pictures, with one dialogue box below each pair to enter what features participants perceived as different between the two pictures. Participants received 60 seconds for each pair, so the distractor task lasted 180 seconds.

Phase II contained two factors: 2 (intervening task: test, restudy) × 2 (intervening task processing: relational, item-specific). In the test with item-specific processing condition, participants saw cue words together with the first letters of the corresponding target words. They were asked to complete and fill in the target words as they

remembered. Six sets with six cue-target word pairs in each set were created. No target word in the same semantic category was in the same set. The word pairs were pseudo-randomized within and between sets. In each set, there were 40 seconds allocated to the test and 20 seconds for feedback. The only difference between the test with relational processing condition and the previous condition was that the six sets were grounded in semantic categories of target words. The reason why first letters of target words were provided was to ensure performance was at ceiling in the intervening tests, to fully capture the testing effect. Restudy conditions differed from test conditions in that target words were presented in full, instead of just the first letters. Participants received 60 seconds to read the six cue-target pairs in each set. Phase II and Phase III were separated by another distractor task with pictures different from the previous one.

Phase III consisted of final test processing (relational, item-specific). In free recall, there was a dialogue box for participants to retrieve and input target words from prior learning. In cued recall, participants were asked to type the target words in response to cue words. The cue words were presented randomly with each on a separate page. Participants did not have a time limit for either type of recall test.

# Chapter 3: Results

## Intervening Test Performance

During the intervening task, participants in the test condition underwent an extra test, because no response was recorded in the restudy condition. Average percentages of target words correctly recalled are presented in *Figure 2*. It can be observed that averaged across intervening task processing, performance was higher for participants presented with a test tapping relational processing ($M$ = 96%) than those presented with a test calling for item-specific processing ($M$ = 90%). An ANOVA with intervening task processing (relational, item-specific) and final test processing (relational, item-specific) as two between-subject factors revealed that there was a significant main effect of the type of cognitive processes the intervening task triggered ($F(1, 71) = 14.287$, $p < .001$, $\eta^2 = .162$).

## Final Test Performance

Mean percentages of target words correctly recalled during the final test in each condition are illustrated in *Figure 3*. According to descriptive statistics, it was obvious that as intervening tasks, the test condition improved performance in the final test more than restudy in every combination of processing. Also evident was that an intervening relational task ($M$ = 71%) led to higher performance than an intervening item-specific task ($M$ = 60%) and a final item-specific test ($M$ = 73%) resulted in more target words correctly recalled than a final relational test ($M$ = 59%). One interesting finding was that taking a test appeared to be more advantageous than restudying in terms of final test

performance when the final test called for item-specific processing, as indicated by the mean percentages (IS-R: 55% vs. 52%; R-R: 66% vs. 62%; IS-IS: 74% vs. 61%; R-IS: 84% vs. 72%).

To test if the above comparisons were statistically significant, a three-way between-subject ANOVA with intervening task (test, restudy), intervening task processing (relational, item-specific), and final test processing (relational, item-specific) as variables was conducted. Three main effects of test over restudy during the intervening task ($F(1, 143) = 5.127$, $p = .025$, $\eta^2 = .029$), relational over item-specific processing during the intervening task ($F(1, 143) = 9.485$, $p = .002$, $\eta^2 = .054$), as well as item-specific over relational processing in the final test ($F(1, 143) = 16.347$, $p < .001$, $\eta^2 = .093$) were found. Although there appeared to be a difference between test and restudy when the final test processing was item-specific than when it was relational, there was no significant interaction between intervening task and final test processing ($F(1, 143) = 1.667$, $p = .199$, $\eta^2 = .009$). Inconsistent with this study's prediction of TAP, it was highly unlikely that there was an interaction between cognitive processes used in the intervening task and the final test ($F(1, 143) < 1$, $p = .995$, $\eta^2 < .001$).

Given that average recall accuracy on the intervening test was not the same between the relational test and the item-specific test, which may have affected final test performance, conditioned final test recall accuracy was calculated and analyzed. This was done by excluding data for target words not retrieved in the intervening test and computing percentage correctly recalled in the final test as a function of target words correctly retrieved in the intervening test. Details of the descriptive statistics of each

condition are reported in *Figure 4*. There was little difference between corrected and raw recall accuracy. For example, means of percentage correctly recalled changed from 74% to 76% in test-IS-IS, from 84% to 85% in test-R-IS, and from 66% to 65% in test-R-R. Results revealed by a new three-way ANOVA remained the same. There were three main effects (intervening task: $F(1, 143) = 6.373$, $p = .013$, $\eta^2 = .035$; intervening task processing: $F(1, 143) = 9.048$, $p = .003$, $\eta^2 = .050$; final test processing: $F(1, 143) = 19.109$, $p < .001$, $\eta^2 = .106$) and no interaction implying transfer of processing from the intervening task to the final task ($F(1, 143) = .010$, $p = .920$, $\eta^2 < .001$).

### MEASURES OF RELATIONAL PROCESSING IN FREE RECALL

A series of measures were conducted on all final free recall responses to take a closer look at involved cognitive processes, especially relational processing, to validate the assumption that the intervening tasks did affect relational and item-specific processing. These measures included category clustering, number of categories correctly recalled, and number of items per category correctly recalled. Category clustering is an appropriate criterion for relational processing, because it estimates the extent to which words are organized based on similarities in a free recall response. In comparison, number of categories recalled and number of items per category recalled are less pure assessments of relational processing. For example, number of items per category recalled was used as an inter-item relational processing indicator by Mulligan and Peterson (2015), but as an item-specific processing indicator by Rawson and colleagues (2015) and other earlier studies (e.g., Hunt & Seta, 1984). It is reasonable that the more

organized a response is, the more words per category it includes. However, an almost perfect recall with little clustering can still be possible, suggesting that number of items per category recalled does not solely measure relational processing. Category clustering was calculated based on the adjusted ratio of clustering (ARC) index (Roenker, Thompson, & Brown, 1971). According to the scale, a score of 0 indicates no clustering, whereas a score of 1 is equivalent to perfect clustering. Each free recall response was coded and entered into an established ARC score calculator (Senkova & Otani, 2012). Calculation of number of categories recalled and number of items per category recalled were more straightforward.

**Adjusted Ratio of Clustering**

Means and standard errors of ARC scores of conditions with a relational final test are plotted in *Figure 5*. Visually, relational processing ($M$ = .79) during the intervening task led to higher category clustering in the final free recall than item-specific processing ($M$ = .48). In a follow-up ANOVA with two factors intervening task and intervening task processing, this comparison was significant ($F(1, 72) = 14.771$, $p < .001$, $\eta^2 = .168$), indicating that relational processing during the intervening task helped participants recognize that target words could be organized in their semantic categories.

**Number of Categories Correctly Recalled**

Descriptive statistics of number of categories recalled are noted in *Figure 6*. In all the conditions combining type and processing of the intervening task, participants recalled around 5.50 categories in the final test. There was no difference among the

conditions in number of categories recalled (intervening task: $F(1, 72) = .201$, $p = .655$, $\eta^2 = .003$; intervening task processing: $F(1, 72) = .558$, $p = .457$, $\eta^2 = .008$). Retrieval practice or relational processing during the intervening task did not make any difference.

**Number of Items per Category Correctly Recalled**

Average numbers of items per category correctly recalled are illustrated in *Figure 7*. The pattern is similar to that of *Figure 5* in that participants in the relational intervening task ($M = 4.22$) correctly recalled more words in each category in the relational final test than those in the item-specific intervening task ($M = 3.44$). A 2 × 2 between-subject ANOVA supported this statement about intervening task processing ($F(1, 72) = 6.445$, $p = .013$, $\eta^2 = .082$). The intervening task that grouped cue-target associations regarding semantic categories of the target words also allowed participants to remember more words per category.

In summary, compared to restudy, retrieval practice did not significantly enhance clustering target words in their related semantic categories, neither did it elevate number of categories and number of items in each category correctly recalled. The only thing during the intervening task that mattered was the type of cognitive processes participants engaged with. There was more relational processing in the final task after the relational intervening task than after the item-specific intervening task, as indicated by both category clustering and number of target words per category being significantly higher in the former case. At least based on these results, TAP for relational processing was supported.

### TEST VS. RESTUDY IN RELATIONAL INTERVENING TASK FOLLOWED BY RELATIONAL FINAL TEST

To achieve the second goal of this study, that is, to replicate Experiment 1 by Peterson and Mulligan (2013), the conditions comparing test and restudy under the same type of processing was extracted. Percentage of target words, ARC, number of categories, together with number of target words per category in the final free recall after the relational intervening task are summarized in Table 2. A *t*-test comparing mean percentages of target words correctly recalled between test and restudy showed non-significant results ($t(36) = .442$, $p = .661$, Cohen's $d = .143$). Peterson and Mulligan (2013) found a negative testing effect as implied by significantly higher percentage correct and ARC after restudy than test. This was supplemented by higher number of items recalled per category in a reanalysis by Mulligan and Peterson (2014). This current study did not find significant testing effects on any of these variables.

Also presented in Table 2 were the 5 experiments conducted by Rawson and colleagues (2015). Experiments 1 to 4b repeated the same procedure. What was different from Peterson and Mulligan (2013, Experiment 1) was that participants were asked to type in their responses to the intervening test or read silently during restudy, instead of reading aloud in both conditions. Participants in Experiment 4a and 5 were instructed to actively produce the target words even during restudy. The part of the current study extracted was very close to Experiment 1 to 4b in design, except for presenting target words in the same semantic categories on the same page. Looking at averaged descriptive

statistics of Experiment 1 to 4b (Rawson et al., 2015), there was minimal difference between test and restudy on these proportion correct and relational processing measures.

# Chapter 4: Discussion

In line with the goals of this study to test the TAP account of the testing effect and replicate Experiment 1 conducted by Peterson and Mulligan (2013), the analysis was split into two sections. When all responses to the final test were included, it was found that there was no interaction between the type of processing involved in the intervening task and the final test, a pattern not favoring TAP. This was because the relational intervening task universally improved final test performance compared to the item-specific intervening task. Furthermore, performance on the final test was higher when the final test itself was item-specific instead of relational. The overall effect of relational processing elicited by the corresponding intervening task was not due to slightly higher performance in the relational intervening test, according to the corrected percentages of target words correctly recalled. When relational final test responses were pulled out for further measures of relational processing, the relational intervening task led to more clustering of target words based on their semantic categories. Although evaluated using the relational final test only, it was likely that the relational intervening task improved relational processing more than the item-specific task regardless of the type of the final test.

Two specific conditions representing Peterson and Mulligan's Experiment 1 (2013) were extracted: relational intervening test followed by relational final test and relational restudy followed by relational final test. In terms of percentage of target items correctly recalled, there was no testing effect in the positive direction, neither was a

negative testing effect present. Moreover, there was no significant advantage of test over restudy on the relational processing measure category clustering, even though higher values were obtained for test.

### IS THE TAP ACCOUNT SUPPORTED?

On the surface, this study does not support TAP. To start the discussion, let us assume every learning activity in this study was process pure. That is, in the intervening task where cue-target associations were grouped into semantic categories of target words and presented together, participants only utilized their relational processing capacities; whereas when the associations were shown randomly, relational processing capacities were put aside, because this was an item-specific task. Similarly, free recall and cued recall with rhyming word cues as final tests facilitated relational and item-specific processing respectively. A transfer of processing from retrieval practice to the criterial test appeared when the criterial test was free recall. Here, free recall performance was higher when the intervening task was relational than when it was item-specific. However, when the final test was cued recall that drew on item-specific processing, still those participants who underwent relational processing outperformed those who experienced item-specific processing, which was contrary to the TAP expectation that performance is always preferable when it calls on the same processing that has been previously executed.

This study generated similar results to the three studies contrasting TAP (Glover, 1989; Carpenter & DeLosh, 2006; Kang et al., 2007), since there was one type of retrieval practice that was more beneficial than other types on all criterial measures.

However, not all the theories supported by those three studies (Glover, 1989: encoding variability; Carpenter & DeLosh, 2006: elaborative retrieval; Kang et al., 2007: retrieval difficulty and effort) are suitable to explain the results of this study. Established theories that could interpret this study will be described as follows.

These findings do confirm the elaborative retrieval (Carpenter, 2009) and bifurcation theories (Kornell, Bjork, & Garcia, 2011), at least the part of the bifurcation model dealing with retrieval threshold. The encoding variability account (Martin, 1968; Melton, 1970) is not supported. This is because neither creating a relational task by grouping target words into categories nor paring target words with cue words introduced more unique cues to targets than one another. Theories about retrieval difficulty and effort (Bjork, 1975; Bjork & Bjork, 1992) are also not supported. Participants recalled more target words in the relational intervening test than the item-specific intervening test, implying the latter being more difficult. However, the item-specific intervening task did not lead to more target words recalled in the final test than the relational intervening task. It is not surprising that this study is consistent with what the bifurcation model would predict, since free recall tests, compared to cued recall tests are more difficult and require higher retrieval thresholds. How the elaborative retrieval account is corroborated is evidenced by higher values in category clustering after the relational intervening task. This means that the relational intervening task allowed participants to use taxonomic categories as semantic mediators to organize and remember the target words.

The relational intervening task facilitating the relational final test seems convincing. What is puzzling is that even when the final test was item-specific, the

relational intervening task still led to higher performance than the item-specific one. Does this mean that the relational intervening task enhanced item-specific processing to a greater extent than the item-specific intervening task? Or was the item-specific final test in reality relational in nature? Potentially confounding factors in the relational intervening task may render the first conjecture possible.

**Relational and Item-Specific Processing vs. Item-Specific Processing**

Item-specific processing appeared to be a confounding variable in the relational intervening task and final test. In the item-specific intervening task, target words were paired with their rhyming cues and presented in a random order, whereas in the relational intervening task, cue-target pairs were organized according to semantic categories of the target words and presented as groups. The key message here is that item-specific information was present in the relational intervening task. To be more specific, the item-specific intervening task facilitated item-specific processing only, while the relational intervening task made both relational and item-specific processing possible. A similar situation applies to free recall and cued recall. Free recall requires a considerable amount of both relational and item-specific processing (Einstein & Hunt, 1980; Hunt & Einstein, 1981), thus the intervening task that was superficially only relational, but in reality also item-specific resulted in a better free recall performance. It is known that even though recognition is predominantly driven by item-specific processing, a small amount of relational processing would further increase recognition performance (Einstein & Hunt, 1980; Hunt & Einstein, 1981). When this is applied to cued recall, the intervening task

that promoted gains in both relational and item-specific information naturally lead to more targets words recalled.

If the above assumption is correct, looking at types of processing involved in the intervening task alone, this study is in line with the encoding variability (Martin, 1968; Melton, 1970) and episodic context theories (Karpicke, Lehman, & Aue, 2014). Because performance as measured by percentage of target words, category clustering, and number of items per category recalled was higher when types of retrieval cues were maximized. That is, when the intervening task facilitated both relational and item-specific processing. These relational and item-specific cues then increased retrieval routes to the corresponding targets to a greater extent than those cues in the item-specific only intervening task. The aggregate of relational and item-specific cues then restricted the set of potential target words. For instance, it is not difficult to imagine that there are more potential target words that rhyme with the cue word "*Moon*" than that those that both rhyme with "*Moon*" and belong to the semantic category kitchen utensils. This finding is similar to some existing research. For example, relational and item-specific processing together facilitated maximal recall performance (e.g., Einstein & Hunt, 1980; Hunt & Einstein, 1981). In the same vein, processing weakly related learning materials by creating a structure and processing strongly related learning materials by accentuating uniqueness are sufficient to improve recall (e.g., Huff & Bodner, 2014; McDaniel, Einstein, & Lollis, 1988).

The relational and item-specific tasks in this study as indicated by the results are analogous to free recall, cued recall and recognition in the studies by Glover (1989),

Carpenter and DeLosh (2006), and Kang and colleagues (2007). This is because one type of intervening task appeared to be more beneficial than others to improving performance on all types of final test. As mentioned several times, successful free recall depends on both relational and item-specific information and facilitates both types of processing. It is possible that both types of processing can be enhanced to a larger extent by free recall than by cued recall or recognition. It is highly likely that free recall, cued recall, and recognition are just nominally different. In terms of underlying cognitive processes, they are not truly dissociable and are not suitable for studies investigating TAP.

Above discussion is based on the assumption that the relational intervening task called for item-specific processing, in addition to relational processing. However, this may not be the case. One participant informed the experimenter that once she noticed the target words were shown in categories, she stopped paying attention to the cues. To determine whether the relational intervening task truly involved item-specific processing and the extent to which this item-specific processing was greater or less than item-specific processing elicited by the item-specific intervening task, relatively pure item-specific processing measures need to be acquired. Recognition is a good candidate (Einstein & Hunt, 1980; Huff & Bodner, 2014). Intrusion analysis was not conducted, because wrong answers may originate from different types of processing in different final tasks. To be specific, wrong answers in cued recall likely resulted from finding words that rhyme with the cues (item-specific processing), whereas intrusions in free recall were likely instances in the covered categories (relational processing). If item-specific processing existed in both the relational intervening task and the relational final test, this

study testing TAP needs to be adjusted to accommodate purer processing. This could be done by removing item-specific processing in both phases, such as changing the blocked presentation of rhyming cue-target pairs into blocked presentation of category cue-target pairs. In addition, free recall could be replaced by a category cued recall.

**Semantic Processing vs. Phonological Processing**

Another possible confounding factor was that semantic and phonological processing covaried with relational and item-specific processing across conditions. In the intervening task aiming at encouraging relational processing, cue-target associations were presented in their according semantic categories. In the item-specific intervening task, cue-target associations were presented randomly, but the cues rhyme with the target words. The same explanation applies to the final tests. If this was what actually happened, then the results could be rephrased as: in the intervening task phase, the semantic task had an overall effect on number of target words recalled in the final test in a positive direction; in the final test phase, the phonological task allowed higher number of target words to be recalled.

Still, there was no interaction to support a TAP account of the testing effect. Previously cited studies (McDaniel & Masson, 1985; Morris et al., 1977) also found comparable main effects. McDaniel and Masson's (1985) study noted higher final performance after a semantic intervening test and during a rhyming final test. In comparison, main effects of semantic intervening test and semantic final test were evident in the study by Morris and colleagues (1977), this was because the rhyming final

test was designed more difficult. Interactions were present in these two studies, but absent in the current study. Considering semantic and phonological processing alone, this study supports the level of processing account (Craik & Lockhart, 1972; Craik & Tulving, 1975). This is because semantic processing is deeper and more elaborative (Anderson & Reder, 1979). Moreover, semantic information is used as a media to constitute meaning in our daily lives (Floridi, 2005). Note that this study is not in favor of difficulty and strength theories of the testing effect (Bjork, 1975; Bjork & Bjork, 1992). In this study, the semantic intervening task was actually easier than the phonological intervening task, as implied by extra category information provided and higher intervening test performance. Echoing with one of the limitations of difficulty and strength theories, this explanation highlights that the term "difficulty" is indeed vague. The theories themselves originated from the level of processing account (Craik & Lockhart, 1972; Craik & Tulving, 1975), potentially because leaning activities at a deeper level are usually more difficult. However, this is not always the case. Therefore, "difficulty" is an intuitive term, but not an accurate description.

The semantic and phonological processing layer may have interacted with the relational and item-specific processing layer and further complicated this study. A precedent is the study by Einstein and Hunt (1980). Aside from their main goal to compare relational and item-specific processing, they inserted different levels of semantic and phonological processing into each level. When relational and item-specific tasks were compared during purely semantic or phonological processing, relational and item-specific processing had proportional influence to performance. Specifically, there was no

difference between taxonomic categorizing (relational/semantic) and pleasantness rating (item-specific/semantic) on final free recall. The latter contributed more to final recognition, whereas the former helped with category clustering in final free recall. When the comparison was between first letter categorizing (relational/phonological) and rhyme rating (item-specific/phonological), a similar pattern emerged except that category clustering was consistently at chance level. However, when taxonomic categorizing (relational/semantic) and rhyme rating (item-specific/phonological) were compared against each other, it was found that the relational task was more beneficial than the item-specific task on all measures of final test: free recall, category clustering in free recall, as well as recognition, suggesting that there was a significant amount of additional benefit introduced by the semantic portion of the relational task. This was likely what happened in the current study.

To remove this confounding variable and truly dissociate relational and item-specific processing, the relational and item-specific tasks need to be made purely semantic or purely phonological. Creating a phonological relational task is difficult and using first letter categorizing like Einstein and Hunt (1980) may result in orthographic processing confounding with phonological processing. Another possibility is to change the existing phonological item-specific task into a semantic item-specific task. This is an economic option, because semantic cues of the same target words used in this study have already been created and experimented by Rawson and colleagues (2015). Furthermore, this will open up more replication opportunities.

## Is Experiment 1 By Peterson And Mulligan (2013) Replicated?

The answer is a tentative no. Experiment 1 by Peterson and Mulligan (2013) is not fully replicated. Instead of a negative testing effect on number of target words recalled, test and restudy were not significantly different. In addition, there was no difference between test and restudy on the relational processing measure category clustering, although test led these values into a slightly positive direction. These findings suggest that test did not significantly modify overall performance and relational processing. The reason why this happened may be due to subtle differences in the designs employed in this study and by Peterson and Mulligan (2013), which will be discussed later.

This study yielded very similar results to Experiment 1 to 4b by Rawson and colleagues (2015). Intervening test and restudy resulted in similar values in proportion of target words correctly recalled, category clustering, number of categories and number of items per category correctly recalled in the final test. Note that data for Experiment 1 to 4b were averaged across identical experiments (proportion correct was averaged across Experiment 2 to 4b). Therefore, no inferential statistics were available. There were only three significant comparisons: test compared to restudy led to higher proportion correct in Experiment 3, as well as higher number of items per category in Experiment 1 and 3, indicating that overall the difference in processing between test and restudy was actually minimal. These similarities in results likely originated from similarities in designs. During the intervening task both studies asked participants in the test condition to type in their responses, whereas there was no overt response in the restudy condition. The reason

why ARC scores were higher in this study was probably because blocked cue-target associations were presented on the same page, compared to one pair on each page, thus made relational information more salient.

Relational processing was not enhanced more by taking the intervening test than restudying. What about item-specific processing? The thesis of the research by Peterson and Mulligan (2013), as well as Mulligan and Peterson (2014) is that while the intervening test enhanced item-specific and intra-item relational processing (item-specific in the context of this study), inter-item relational processing (relational processing in the context of this study) is compromised. If greater item-specific processing could be demonstrated in this study, the multifactor account (Mulligan & Peterson, 2014; Peterson & Mulligan, 2013) could at least be partially corroborated. Existing data is not sufficient to provide a decisive answer. Looking at all the proportion correct data, it does appear that when the final test was item-specific, intervening test was far greater than restudy, whereas test and restudy led to similar performance when the final test was relational. This pattern suggests that there may be special mechanisms involved in test that enhanced item-specific processing. However, there was no significant interaction between intervening task and final test processing. The reason why this non-significance occurred may be that cued recall did not provide a pure measure of item-specific processing. A future study that aims to compare test against restudy on item-specific processing could employ a recognition final test (Einstain & Hunt, 1980; Huff & Bodner, 2014), similar to Experiment 2 of Mulligan and Peterson (2014).

When all the experiments replicating Experiment 1 by Peterson and Mulligan (2013) are put together, it is obvious that subtle differences in designs led to testing effects in completely different directions. Peterson and Mulligan (2013, Experiment 1) found a strong negative testing effect in proportion of target words correct, category clustering, and number of target words per category. Using the exact same procedure, a marginally negative testing effect was only identified in category clustering (Rawson et al., 2015; Experiment 5). There was no testing effect obtained in this study or Experiment 1 to 4b by Rawson and colleagues (2015). The only incident of positive testing effect was found in Experiment 4a (Rawson et al., 2015) where test was more beneficial than restudy on proportion of target words, category clustering, number of categories, and number of target words per category. When the detailed designs are scrutinized, it seems possible that differences in results were driven by different levels of production involved in the intervening tasks. Peterson and Mulligan (2013), as well as Rawson and colleagues (2015, Experiment 5) asked participants to respond to the cues verbally or read aloud the cue-target pairs. In the current study and Experiment 1 to 4b (Rawson et al., 2015), participants typed the target words using computer keyboards in the test condition. In the study condition, there was no explicit response. Experiment 4a (Rawson et al., 2015) was similar to Experiment 1 to 4b, but overt typing of target words was required in the study condition.

This description shows that reading aloud, reading silently, and typing involved in restudy produced increasingly negative restudy effect and increasingly positive testing effect. This pattern can possibly be explained with reference to the dual route model of

visual word recognition and reading aloud (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001), which is a probabilistic model detailing cognitive processes involved from print to speech. When reading aloud, information of words goes through the lexical semantic route. That is, from orthographic input, via the semantic system, to phonological output. Reading silently is different in that less information passes through the lexical semantic route from orthographic input to phonological output. The lexical semantic route may be further inhibited by typing, because typing is heavily reliant on orthographic information. Therefore, reading aloud, reading silently, and typing during restudy may encourage an increasing amount of semantic processing. Suggested as a confounding variable in the current study, semantic processing may have had a profound effect on elaborating the target words and connecting them to their categories.

## CONCLUSION

If all intervening tasks and final tests are regarded as process pure, the TAP account is not supported by the current study. Instead, the relational intervening task demonstrated an overall improvement in retention, regardless of what the final test was. Two potential confounding variables were identified in the relational intervening task and final free recall. They were additional item-specific processing and semantic processing respectively. When the confounding variables are considered, the TAP account is favored. This is because the "relational" intervening task facilitated exactly what final free recall asked for: quantitatively and qualitatively different relational and item-specific processing compared to the item-specific intervening task. It is important to point out that

TAP entails consistency in processing, instead of activities. When the highest performance is not found when encoding/intervening task and final test are matched, TAP would suggest that the tasks are not matched in terms of specific cognitive processes involved. This reasoning, as well as a number of previous studies led to a conclusion that free recall, cued recall, and recognition overlap in processing and cannot equally contribute to the interaction pattern of TAP. To demonstrate the interaction pattern of TAP, more process pure tasks and measures need to be employed.

Unlike Peterson and Mulligan's Experiment 1 (2013) where a negative testing effect was found, test and restudy as intervening tasks were not significantly different on how they affected relational processing and final test performance. The results were very similar to the average results of Experiment 1 to 4b (Rawson et al., 2015). After comparing a series of experiments replicating Experiment 1 of Peterson and Mulligan (2013), it was found that the direction of testing effect changed with the type of learning activities. Reading aloud, reading silently, and typing during restudy gradually shifted the testing effect from negative to positive. Given that these activities vary in the amount of semantic processing involved, it is possible that meaningful and organizational processing are what makes retrieval practice an efficient memory facilitator.

Table 1: Major Theories of Retrieval Practice, Proponents (in Parentheses), Focus on Description or Mechanism, Key Information (in Parentheses), and Phenomena Explained.

| Theory | Cue-target association strength or cognitive processing mechanism? | Phenomena explained |
|---|---|---|
| **Retrieval difficulty** (Bjork, 1975) | Association strength | Difficulty of intervening test |
| **Retrieval strength** (Bjork & Bjork, 1992) | Association strength (Storage strength and retrieval strength) | Difficulty of intervening test |
| **Bifurcation** (Kornell, Bjork, & Garcia, 2011) | Association strength (Encoding strength and retrieval strength) | Difficulty of intervening test and final test |
| **Encoding variability** (Martin, 1968; Melton, 1970) | Processing mechanism (Target remains the same, perceived cues vary; aggregate of cues uniquely point to target) | Completeness of intervening test |
| **Elaborative retrieval** (Carpenter, 2009) | Processing mechanism (Cue and target remain the same, elaborative routes vary) | Difficulty of intervening test |
| **Episodic context** (Karpicke, Lehman, & Aue, 2014) | Processing mechanism (Target remains the same, contextual cues vary; aggregate of cues uniquely point to target) | Difficulty and completeness of intervening test |
| **Transfer-appropriate processing** (Morris, Bransford, & Franks, 1977) | Processing (no mechanism) | Match between intervening test and final test |

Table 2: Comparison among the Current Study, Experiment 1 by Peterson and Mulligan (2013), and Experiment 1-5 by Rawson, Wissman, and Vaughn (2015) on Percentage Correct and Relational Processing Measures.

| Experiment | Proportion Correct | | ARC | | No. of Categories | | No. of Items per Category | |
|---|---|---|---|---|---|---|---|---|
| | Test | Restudy | Test | Restudy | Test | Restudy | Test | Restudy |
| The current study R-R conditions | .66 (.05) | .62 (.06) | .84 (.06) | .73 (.07) | 5.42 (.19) | 5.42 (.19) | 4.35 (.27) | 4.09 (.35) |
| Peterson & Mulligan (2013) Experiment 1 | .45 | .58** | .28 | .49** | 5.18 | 5.43 | 3.14 | 3.73** |
| Rawson et al. (2015) Experiment 1-4b | .64 (.03) | .61 (.03) | .23 (.07) | .26 (.09) | 4.9 (.2) | 4.5 (.2) | 2.7 (.1) | 2.3 (.2) |
| Rawson et al. (2015) Experiment 4a | .65 (.03) | .55** (.03) | .21 (.06) | .05* (.09) | 5.0 (.2) | 4.6* (.2) | 2.5 (.1) | 2.2** (.2) |
| Rawson et al. (2015) Experiment 5 | .60 (.03) | .67 (.03) | .22 (.07) | .37* (.08) | 4.7 (.2) | 4.8 (.2) | 2.6 (.1) | 2.6 (.2) |

Note: Standard errors are reported in parentheses. Decimal places are formatted according to the original studies. All experiments compared test and restudy when the intervening tasks and final tests both required relational processing. Participants were asked to read aloud during test and restudy (Peterson & Mulligan, 2013, Experiment 1; Rawson et al., 2015, Experiment 5), type during test only (Rawson et al., 2015, Experiments 1-4b), type during test and restudy (Rawson et al., 2015, Experiment 4a). The row corresponding to Experiments 1-4b (Rawson et al., 2015) contains values averaged across 4 experiments. Percentage correct in this row was averaged across Experiments 2-4b. Underlined values are statistically higher than the other values in the same pairs. *$p < .10$. **$p < .05$.

Figure 1: Flow Chart Illustrating 8 Between-Subject Conditions Implemented in 3 Phases.

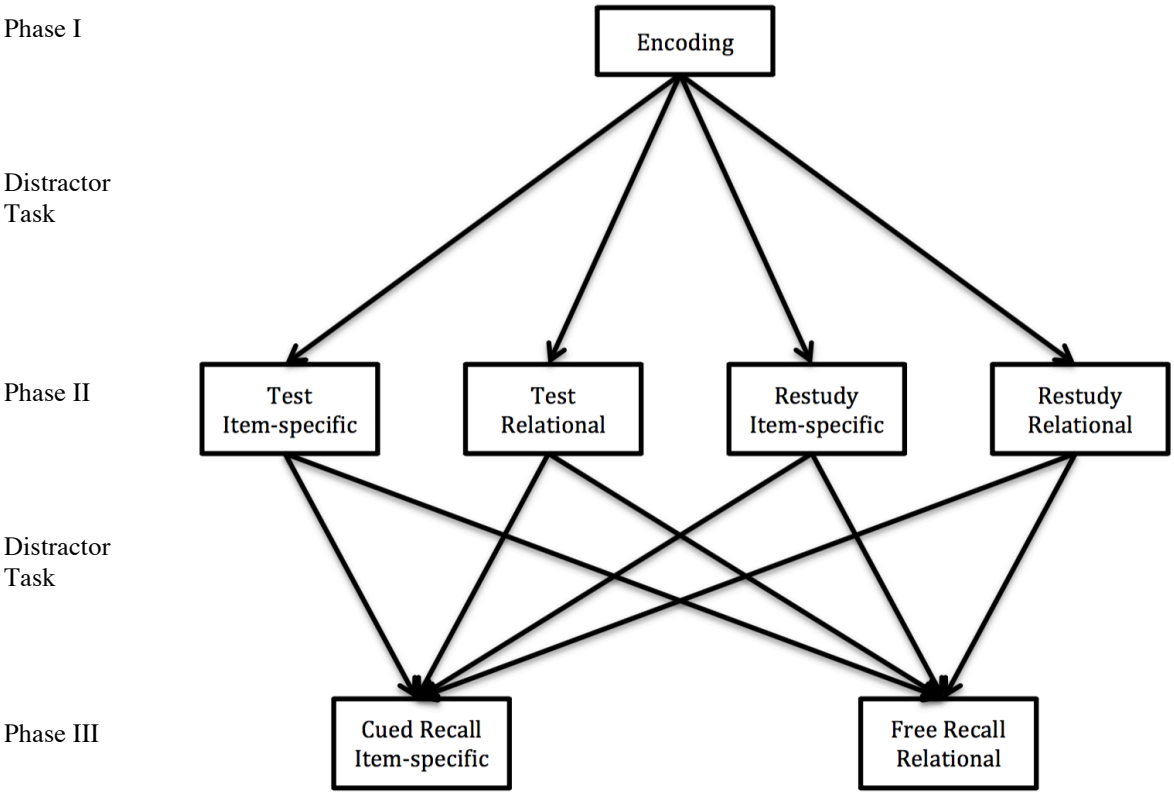


Figure 2: Average Percentage of Target Items Correctly Recalled in the Intervening Test in each Intervening Task Processing × Final Test Processing Combination. Error Bars Indicate Standard Errors.

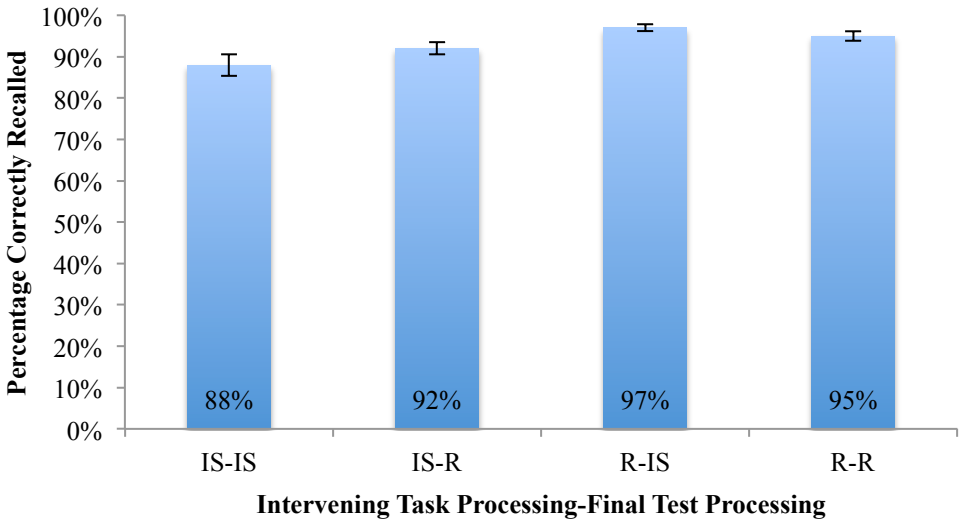


Figure 3: Average Percentage of Target Items Correctly Recalled in the Final Test in each Intervening Task × Intervening Task Processing × Final Test Processing Combination. Error Bars Indicate Standard Errors.

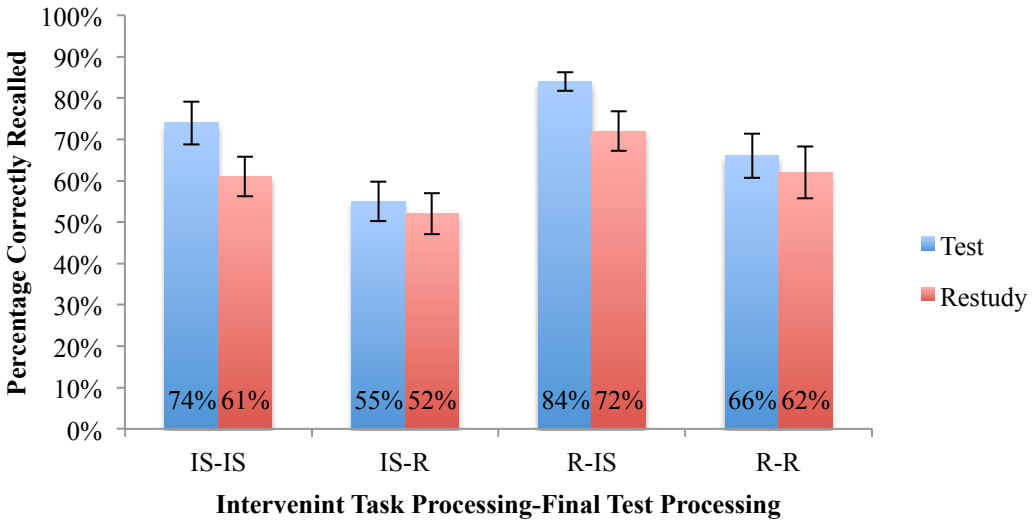


Figure 4: Average Corrected Percentage of Target Items Correctly Recalled in the Final Test in the Intervening Test in each Intervening Task × Intervening Task Processing × Final Test Processing Combination. Error Bars Indicate Standard Errors.

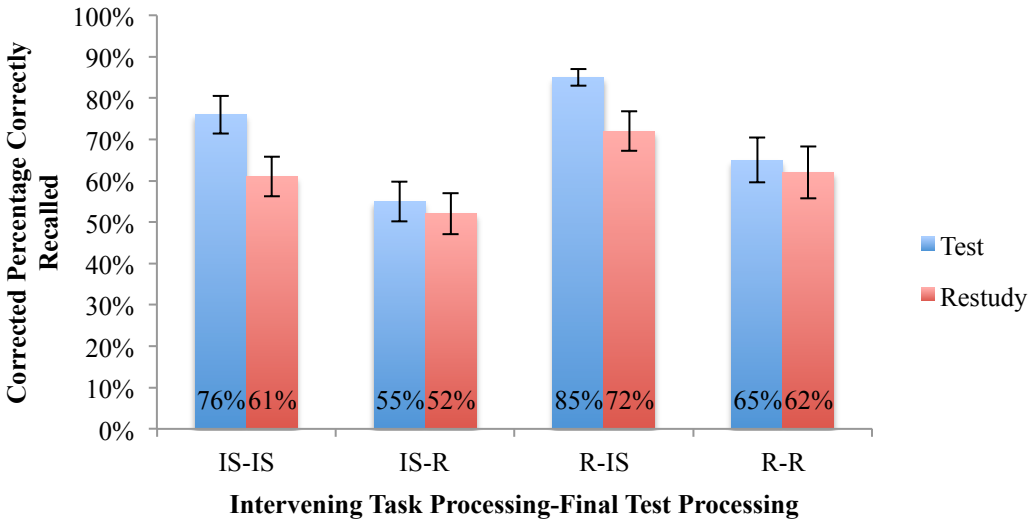


Figure 5: Adjusted Ratio of Clustering in Conditions with a Free Recall Final Test. Error Bars Indicate Standard Errors.

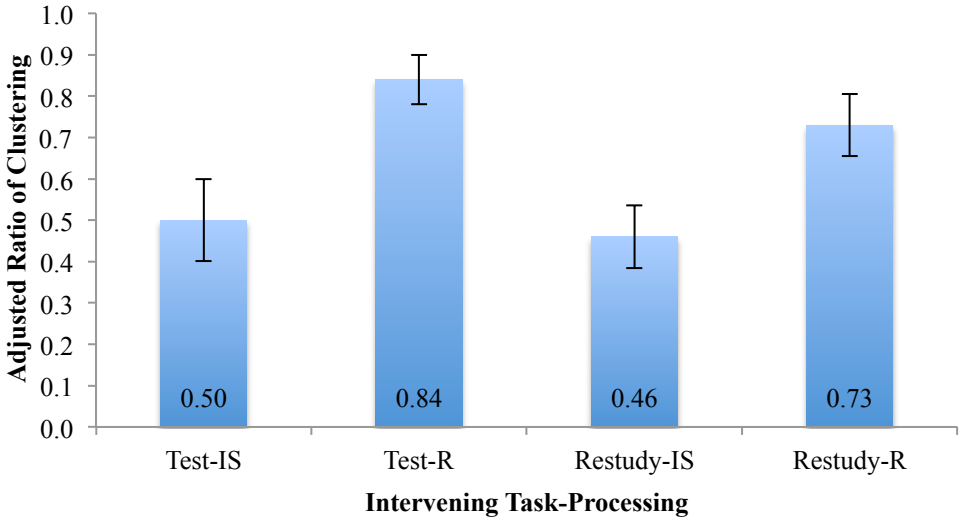


Figure 6: Number of Semantic Categories Recalled in Conditions with a Free Recall Final Test. Error Bars Indicate Standard Errors.

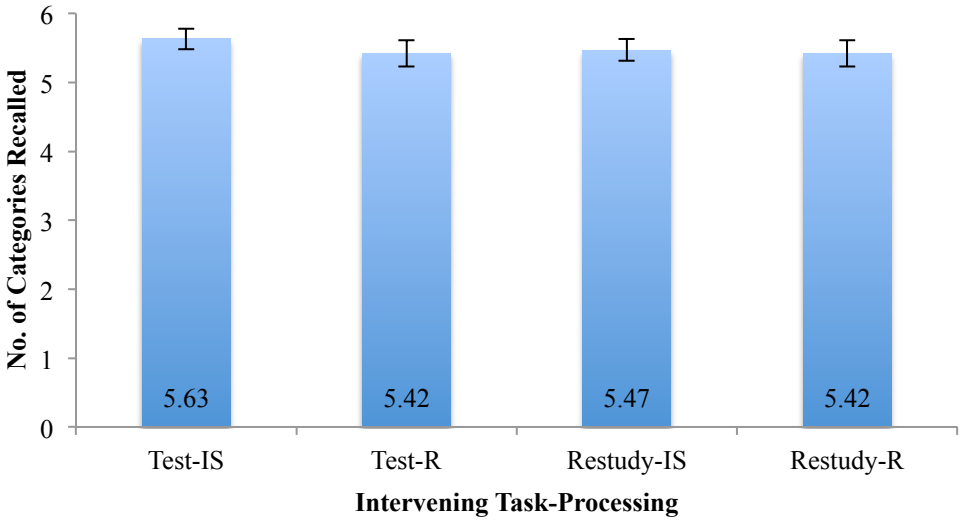
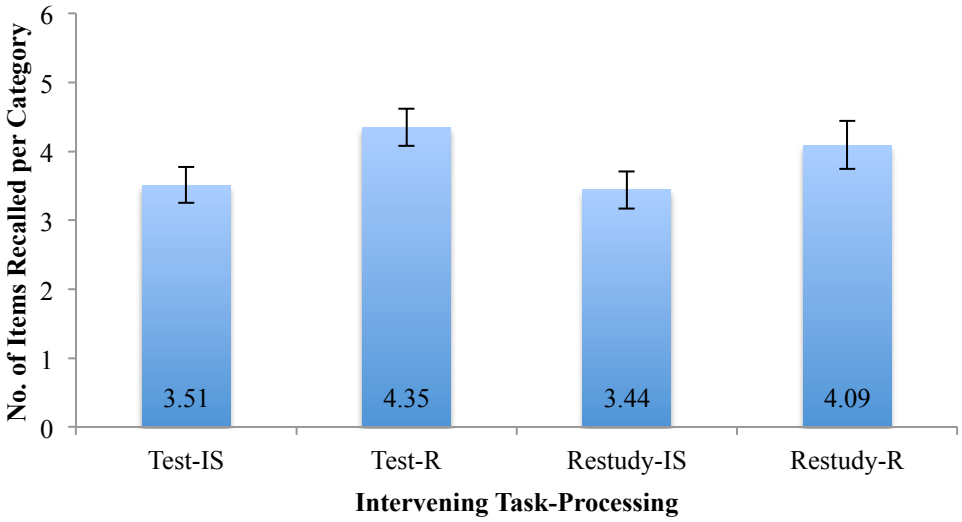


Figure 7: Number of Items per Semantic Category Recalled in Conditions with a Free Recall Final Test. Error Bars indicate Standard Errors.



## Appendix A

## Cue-Target Pairs Modified from Peterson and Mulligan (2013)

| Cue | Target |
|---|---|
| Category: A four-footed animal | |
| Force | Horse |
| Swear | Bear |
| Vow | Cow |
| Cheer | Deer |
| Blouse | Mouse |
| Rig | Pig |
| Category: A fruit | |
| Tape | Grape |
| Fair | Pear |
| Teach | Peach |
| Drum | Plum |
| Time | Lime |
| Tune | Prune |
| Category: A kitchen utensil | |
| Wife | Knife |
| Cork | Fork |
| Moon | Spoon |
| Ban | Pan |
| Disk | Whisk |
| Skate | Plate |
| Category: A metal | |
| Feel | Steel |
| Cold | Gold |
| Win | Tin |
| Pickle | Nickel |
| Class | Brass |
| Think | Zinc |
| Category: A part of the human body | |
| Beg | Leg |
| Linger | Finger |
| Bread | Head |
| Doe | Toe |
| Sand | Hand |
| Hose | Nose |

## Appendix A (Cont.)

| Cue | Target |
|---|---|
| Category: A transportation vehicle | |
| Jar | Car |
| Plus | Bus |
| Puck | Truck |
| Cane | Plane |
| Hike | Bike |
| Coat | Boat |

## References


Anderson, J. R., & Reder, L. M. (1979). An elaborative processing explanation of depth of processing. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of Processing in Human Memory* (pp. 385-404). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information Processing and Cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

Bjork, R. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition* (pp. 185–205). Cambridge, MA: The MIT Press.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From Learning Processes to Cognitive Processes: Essays in Honor of William K Estes: Vol. 2*. (pp. 35–67). Hillsdale, NJ: Erlbaum.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: the benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1563-1569.

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(6), 1547-1552.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory and Cognition*, *34*(2), 268-276.

Collins, A. M., & Quillian, M. R. (1972). Experiments on semantic memory and language comprehension. In L. Gregg (Ed.), *Cognition and learning* (pp. 117–138). New York, NY: Wiley.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*(1), 204-256.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Bbehavior*, *11*(6), 671-684.

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268-294.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4-58.

Einstein, G. O., & Hunt, R. R. (1980). Levels of processing and organization: Additive effects of individual-item and relational processing. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(5), 588-598.

Floridi, L. (2005). Is semantic information meaningful data? *Philosophy and Phenomenological Research*, *70*(2), 351-370.

Gardiner, F. M., Craik, F. I., & Bleasdale, F. A. (1973). Retrieval difficulty and subsequent recall. *Memory and Cognition*, *1*(3), 213-216.

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory and Cognition*, *7*(2), 95-112.

Glover, J. A. (1989). The" testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*(3), 392-399.

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(4), 801-812.

Huff, M. J., & Bodner, G. E. (2014). All varieties of encoding variability are not created equal: Separating variable processing from variable tasks. *Journal of Memory and Language*, *73*, 43-58.

Hunt, R. R., & Einstein, G. O. (1981). Relational and item-specific information in memory. *Journal of Verbal Learning and Verbal Behavior*, *20*(5), 497-514.

Hunt, R. R., & Seta, C. E. (1984). Category size effects in recall: The roles of relational and individual item information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(3), 454-464.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513-541.

Kang, S. H., McDermott, K. B., & Roediger III, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4-5), 528-558.

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In B. H. Ross (Ed.), *Psychology of Learning and Motivation, Vol. 61* (pp. 237-284). San Diego, CA: Elsevier Academic Press.

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*(2), 85-97.

Martin, E. (1968). Stimulus meaningfulness and paired-associate transfer: an encoding variability hypothesis. *Psychological Review*, *75*(5), 421-441.

McDaniel, M. A., & Masson, M. E. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(2), 371-385.

Mcdaniel, M. A., Einstein, G. O., & Lollis, T. (1988). Qualitative and quantitative considerations in encoding difficulty effects. *Memory and Cognition*,*16*(1), 8-14.

Melton, A. W. (1970). The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behavior*, *9*(5), 596-606.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519-533.

Mulligan, N. W., & Peterson, D. J. (2015). Negative and positive testing effects in terms of item-specific and relational information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 859-871.

Peterson, D. J., & Mulligan, N. W. (2013). The negative testing effect and multifactor account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1287-1293.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437-447.

Rawson, K. A., Wissman, K. T., & Vaughn, K. E. (2015). Does testing impair relational processing? Failed attempts to replicate the negative testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(5), 1326-1336.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20-27.

Roediger, H. L., & Karpicke, J. D. (2006a). Test-enhanced learning taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249-255.

Roediger, H. L., & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181-210.

Roenker, D. L., Thompson, C. P., & Brown, S. C. (1971). Comparison of measures for the estimation of clustering in free recall. *Psychological Bulletin*, *76*(1), 45-48.

Senkova, O., & Otani, H. (2012). Category clustering calculator for free recall. *Advances in Cognitive Psychology*, *8*(4), 292-295.

Tulving, E. (1985). *Elements of Episodic Memory*. Oxford, UK: Oxford University Press.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*(5), 352-373.

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the norms. *Journal of Memory and Language*, *50*(3), 289-335.

Veltre, M. T., Cho, K. W., & Neely, J. H. (2015). Transfer-appropriate processing in the testing effect. *Memory*, *23*(8), 1229-1237.

Whitten, W. B. (1978). Initial-retrieval "depth" and the negative recency effect. *Memory and Cognition*, *6*(6), 590-598.